1 Introduction

Efforts to build a predictive model for IMDB scores using the 'IMDB_data_Fall_2023' dataset are reported here. The goal of the project was to predict IMDB scores for 12 upcoming movies in the test set scheduled for release in November. We began with data preprocessing and exploratory analysis to understand the variables. Feature selection was performed to identify the most significant predictors. From there, we employed linear regression techniques to build models.

Upon applying our optimal model to the test set, some predictions returned implausible IMDB scores above 10. Further analysis revealed movie budgets in the test set reached extremes beyond the training data bounds. As a result, we were unable to directly apply the model and obtained unrealistic predictions. To address this, we introduced upper bounds for the standardized movie budget variable to constrain predictions to a feasible range. In this report, we detail our full methodology and key results from each modeling step. We also reflect on limitations encountered in directly applying the model to out-of-sample test data with budget outliers. Areas for potential improvement are discussed, such as transforming variables like movie budget to better account for differences between the training and test distributions.

2 Data Description

The 'IMDB_data_Fall_2023' dataset provided consists of 1930 observations with 42 columns. As we are interested in predicting the IMDB score, the dependent variable is the 'imdb_score'.

2.1 Dependent Variable

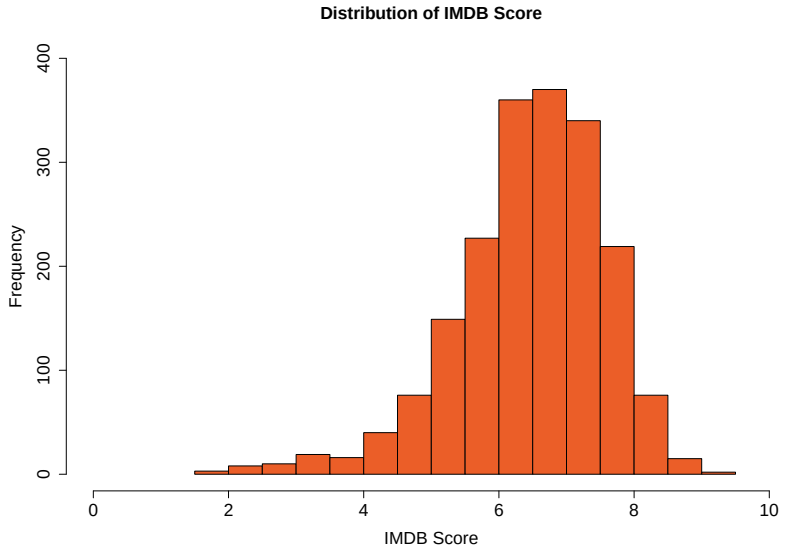

Figure 1: Distribution of IMDB Score

The variable IMDB score has $\mu = 6.511813, \sigma = 1.100033,$ skewness $= 0.865168$. The distribution shows that IMDB scores are slightly left-skewed.

2.2 Independent Variables

2.2.1 Continuous Variables

Table 1: Summary of Statistics of Continuous Variables

Variable	**Mean**	**SD**	**Skewness**
Movie Budget	20973774	14628111	0.5383648
Number of News Articles	770.6093	1860.425	18.62334
Ranking of Main Actor	21189.56	285993.4	23.28595
Ranking of Second Main Actor	17114.09	173281.6	27.59892
Ranking of Third Main Actor	35468.58	260425.8	16.30241
Number of Faces on Poster	1.439896	2.060204	3.73738
IMDBPro Movie Ranking	11612.38	40181.21	14.54413

Most of the categorical variables are mostly skewed right (positive skew values). Only Movie Budget and Number of Faces on Poster have smaller magnitude of skewness.

Within the independent variables, there are a total of 7 continuous variables. For the 7 continuous variables, summary statistics were calculated. See Table 1 for more details. The differing scales across the different variables - for instance budgets measured in millions versus dummy variables on a 0/1 scale - informed the decision to standardize the continuous variables.

2.2.2 Categorical Variables

Table 2: Categorical Variables with High Unique Value Counts

Variable	**Unique Count**
Distributor	334
Director	1115
Plot Keywords	1930
Cinematogropher	737
Production Company	768

As part of the analysis, several variables were identified to have an extremely high number of unique values. The large number of unique values may increase computational costs and degrade model performance (overfitting, instability and difficulty in interpreting results). Table 2 summarizes the variables with high unique counts which will be subsequently dropped for our analysis.

The following variables were explored using factors like frequency, distribution, skew, number of levels and comparison with that of the test set:

- **Release Month**: The data is generally evenly distributed with no changes required.
- **Release Day**: The data is generally evenly distributed with no changes required.
- **Language**: Highly imbalanced with 98% of the movies in the training dataset being English.
- **Country**: Country of production is heavily skewed towards USA (80%) and UK (9%).

- **Maturity Rating**: PG, PG-13, R account for 95% of all ratings.
- **Aspect Ratio**: Aspect ratios 2.35 and 1.85 accounts for 95% of all observations.
- **Colour Film**: Highly imbalanced with 97% of the movies in colour.

2.3 Data Preprocessing

The following considerations were taken to process the data before modelling the prediction model:

- **Standardization of Continous Variables**: While min-max scaling offers benefits with regard to categorical data, we were concerned about outliers. As such, Z-score standardization was employed to place all numeric predictors on a comparable centralized scale, allowing better study of the relationship between the variables.
- **Reducing Noise in the dataset:** To reduce noise in the dataset, observations that were not the majority proportion in columns 'Language', 'Maturity Rating' and 'Aspect Ratio' were dropped.
- **Introducing 'Day of Week**': Both 'Release Day' and 'Release Month' were generally uniformly distributed and were deemed to lack discriminative power on their own as sole categorical predictor. As such, the 'Release Day', 'Release Month' and 'Release Year' were combined into a composite 'Day of Week. variable. This encoding of the release schedule at a weekly resolution was expected to capture meaningful patterns in film debut schedules.
- **Year:** To avoid an excessive number of dummy variables as the 'Year' variable ranges from 1936 - 2018, we treated 'Year' as a continuous variable in this analysis.
- **Dummifying the categorical variables**: Dummies were created for each categorical variable using a hot encoder.
- **Outlier removal:** Outliers were removed by conducting the Bonferroni outlier test and visual analysis throughout our model selection stage.
- **Multicollinearity Check**: A Variance Inflation (VIF) test was performed on the new dataset. Some of the Maturity Rating dummy variables - specifically 'maturity_ratingPG', 'maturity_ratingPG-13' and 'maturity_ratingR' - had unexpectedly high VIF scores (significantly larger than 4). See Table 6 in the Appendix for more details. These predictors were had high linear correlations with other predictors and were dropped.

3 Model Selection

3.1 Methodology

First, we studied the scatterplots of individual predictors against IMDB score and built 29 univariate linear regression models in the form of:

$$\textbf{imdb\_score} = \beta_0 + \beta_1 X_i \textbf{ , where i corresponds to each independent variable post data-processing.} \quad (1)$$

Using the p-values, the predictors were then classified into 'Highly Significant', 'Significant' and 'Marginally Significant'. See Table 7 in the Appendix for variables included for each category. Thereafter, all 17 predictors in the 'Highly Significant' category were used to build a multivariate linear regression model to further determine the signifiance of the predictors. The resulting F-statistic exhibited a notably low p-value (see Table 8 in Appendix), highlighting that the collective influence of these predictors improves the model.

3.1.1 Linearity Test

In determining the linearity of the individual predictors used in the initial multivariate model (see Eqn 3 of the Appendix), we applied both visual analysis (see Fig3 in Appendix) and Tukey test to determine the non-linear predictors. Using a threshold of p-values < 0.01, predictors - 'movie_budget', 'duration', 'nb_news' and 'movie_meter' exhibited non-linear relationships (see Table 9 in Appendix). The model was then tested by modelling these predictors as polynomials from degree 1 to degree 5 and applying ANOVA to determine the most statistically significant polynomial degree. With a p-value < 0.001, a second-degree polynomial is statistically more significant. Testing the model using second degree polynomial increased the adjusted R^2 from 34.4% (Table 9 in Appendix) to 41.1% (Table 10 in Appendix).

3.1.2 Evaluating Heteroskedasticity

Applying a Non-Constant Variance Test, we obtained a p-value of 2.22×10^{-16} which is smaller than 0.05, signalling the presence of heteroskedatisticty. Heteroskedatsticity was corrected with the 'coeftest' function.

3.1.3 Exploring Splines

The use of splines to increase model's predictive power was explored visually. However, the use of splines were not suitable for this model.

3.1.4 Variable Importance Feature

Using the 'caret' package in R and setting random seed = 7 , we re-evaluated the predictors that were initally deemed as highly significant. An importance score was generated for each predictor. Predictors with importance score of less than 2 except for 'release_month_Nov' (IMDB prediction are for movies released in November) were excluded in the final model (see Fig 4 in the Appendix).

3.1.5 Interaction Terms

To account for potential movement between variables and avoid over-attribution, interaction terms were added with the following considerations:

- War-centric films often blend dramatic narratives with action-packed sequences to emphasize the personal stories and intense emotions during times of war. Interaction terms added for 'war'-'drama' and 'war'-'action'.
- Fusion of drama with crime adds depth to a narrative, while combining crime with horror enhances suspense. Interaction terms added for 'crime'-'drama' and 'crime'-'horror'.

- Horror films enriched with drama focus on character development and emotional depth. Interaction term added for 'horror'-'drama'.
- Adventure films are characterized by their action sequences, with the thrill often tied to the action elements. Interaction terms added for 'adventure'-'action' and 'adventure'-'thrill'
- A film's recognition on platforms like IMDB may correlate with its media coverage, suggesting that extensive media attention can boost a movie's visibility and acclaim. Interaction term added for 'movie_meter_IMDBpro'-'nb_news_articles'

4 Results and Discussion

4.1 Model Results

The model selected for testing is shown in Eqn 2:

$$\begin{aligned}\text{imdb\_score} = {} & \beta_0 + \beta_1 \text{release\_monthOct} + \beta_2 \text{sport} + \beta_3 \text{war} + \beta_4 \text{war} \times \text{drama} \\ & + \beta_5 \text{war} \times \text{action} + \beta_6 \text{drama} \times \text{crime} + \beta_7 \text{crime} \times \text{horror} + \beta_8 \text{horror} \times \text{drama} \\ & + \beta_9 \text{action} \times \text{adventure} + \beta_{10} \text{nb\_face} + \beta_{11} \text{action} + \beta_{12} \text{adventure} + \beta_{13} \text{horror} \\ & + \beta_{14} \text{drama} + \beta_{15} \text{crime} + \beta_{16} \text{movie\_budget} + \beta_{17} \text{movie\_budget}^2 \\ & + \beta_{18} \text{release\_year} + \beta_{19} \text{duration} + \beta_{20} \text{duration}^2 + \beta_{21} \text{nb\_news\_articles} \\ & + \beta_{22} \text{nb\_news\_articles}^2 + \beta_{23} \text{movie\_meter\_IMDBpro} + \beta_{24} \text{movie\_meter\_IMDBpro}^2 \\ & + \beta_{25} \text{movie\_meter\_IMDBpro} \times \text{nb\_news\_articles} + \beta_{26} \text{release\_monthNov} \end{aligned} \tag{2}$$

Table 3: Summary of Final Model[1]

	Dependent variable:
	IMDB Score
(*Intercept*)	6.071^{***} (0.067)
β_1 (release_monthOct)	0.046 (0.061)
β_2 (sport)	0.182^{**} (0.091)
β_3 (war)	0.562^{*} (0.330)
β_4 (war×drama)	−0.264 (0.330)
β_5 (war×action)	−0.361 (0.222)
β_6 (drama×crime)	-0.291^{***} (0.096)
β_7 (crime×horror)	0.126 (0.268)
β_8 (horror×drama)	-0.458^{***} (0.160)
β_9 (action×adventure)	−0.022 (0.123)
β_{10} (nb_face)	-0.038^{***} (0.009)
β_{11} (action)	-0.275^{***} (0.060)
β_{12} (adventure)	0.028 (0.078)
β_{13} (horror)	-0.267^{***} (0.075)
β_{14} (drama)	0.474^{***} (0.054)
β_{15} (crime)	0.341^{***} (0.078)
β_{16} (movie_budget)	-6.368^{***} (0.902)
β_{17} (movie_budget^2)	3.631^{***} (0.816)
β_{18} (release_year)	-0.137^{***} (0.022)
β_{19} (duration)	11.614^{***} (0.974)
β_{20} (duration^2)	-3.965^{***} (0.839)
β_{21} (nb_news_articles)	-11.320^{***} (4.134)
β_{22} (nb_news_articles^2)	-9.016^{***} (0.835)
β_{23} (movie_meter_IMDBpro)	-55.474^{***} (9.577)
β_{24} (movie_meter_IMDBpro^2)	5.415^{***} (0.854)
β_{25} (movie_meter_IMDBpro×nb_news_articles)	-3.081^{***} (0.547)
β_{26} (release_monthNov)	0.118^{*} (0.071)
Observations	1,802
Log Likelihood	−2,141.342
Akaike Inf. Crit.	4,336.683
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3 summarizes the coefficients for each of the predictors.

The model in Eqn 2 was evaluated using the Leave One Out Cross Validation (LOOCV) for greater accuracy. LOOCV yielded a Mean Squared Error (MSE) of 0.652. On average, our model prediction deviates by approximately 0.652 from actual IMDB scores. This low MSE is crucial considering that IMDB Score ranges from 0 to 10.

4.2 Model Predictions

Table 4: Prediction for IMDB scores without adjustments

	Movie Title	Predicted IMDB score
1	Pencils vs Pixels	5.02
2	The Dirty South	6.31
3	The Marvels	28.20
4	The Holdovers	6.60
5	Next Goal Wins	6.34
6	Thanksgiving	5.61
7	The Hunger Games: The Ballad of Songbirds and Snakes	17.56
8	Trolls Band Together	6.65
9	Leo	5.80
10	Dream Scenario	5.65
11	Wish	16.45
12	Napoleon	9.92

The same pre-processing steps highlighted in the previous sections was applied to the 'test_data_IMDB_Fall_20 Using the values for the predictors outlined in Table 3, the following predictions for IMDB Scores were made.

4.3 Discussion

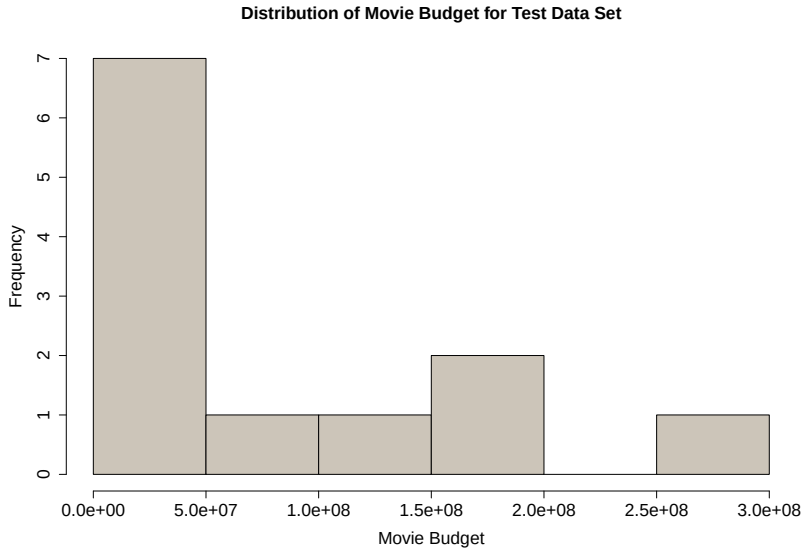

Figure 2: Distribution of Movie Budget for Test Data Set (pre-standardization)

Directly applying our model to the test data resulted in three movies having IMDB Scores of more than 10 (see Table 4): 'The Marvels', 'The Hunger Games: The Ballad of Songbirds', and 'Wish'. Further analysis (Fig 2) showed there were three movies with significantly high budget of more than 150 million dollars. This corresponds to the budget allocated for the three movies highlighted above.

Table 5: Prediction of IMDB scores with adjustments

	Movie Title	Predicted IMDB score
1	Pencils vs Pixels	5.02
2	The Dirty South	6.31
3	The Marvels	6.08
4	The Holdovers	6.60
5	Next Goal Wins	6.34
6	Thanksgiving	5.61
7	The Hunger Games: The Ballad of Songbirds and Snakes	6.93
8	Trolls Band Together	5.83
9	Leo	5.80
10	Dream Scenario	5.65
11	Wish	5.82
12	Napoleon	6.70

While we standardized the movie budget for the test data before running our predictions, our model in Eqn 2 had some limitations in predicting the IMDB scores for movies with high budget. From Table 3, we see that there is a large positive coefficient for β_{17}. This could possibly explain the IMDB scores for the three movies that went above 10. For more details, see Fig 5 for the standardized distribution of Movie Budget for Test Data Set. Recognising the limitations of our model, we implemented a lower and upper bound of z-score for Movie Budget of ± 3. Using a z-score of 3, we account for 99% of the data in the training data set. For the movies highlighted in Table 4, we proceeded on with our predictions using a z-score of 3 for those movies. Resultingly, our final predictions are:

Our final predictions mainly ranges from 5 to 7 which is close to the mean of the IMDB score of the training set (see Fig 1). While we expect movies to receive higher ratings when movie budget increases, on reflection, our model could have over-emphasized the significance of movie budget. On hindsight, we could have studied the ranges of the test data and compare that with the training data to determine what are the possible transformation methods to apply. For example, since some movies in the test data has movie budget that is several folds that of the training set, we could have employed Logarithmic transformation to better account for the relationship and to maintain the feasbility of our predictions.

Appendix

Multivariate Equation:

$$\begin{aligned}\text{imdb\_score} =& \beta_0 + \beta_1 \text{nb\_faces} + \beta_2 \text{action} + \beta_3 \text{adventure} + \beta_4 \text{scifi} + \beta_5 \text{thriller} + \beta_6 \text{horror} + \beta_7 \text{drama} \\ &+ \beta_8 \text{war} + \beta_9 \text{crime} + \beta_{10} \text{movie\_budget} + \beta_{11} \text{release\_year} + \beta_{12} \text{duration} + \beta_{13} \text{nb\_news\_articles} \\ &+ \beta_{14} \text{movie\_meter\_IMDBpro} + \beta_{15} \text{release\_monthDec} + \beta_{16} \text{release\_monthNov} \\ &+ \beta_{17} \text{day\_of\_weekWednesday} \end{aligned} \tag{3}$$

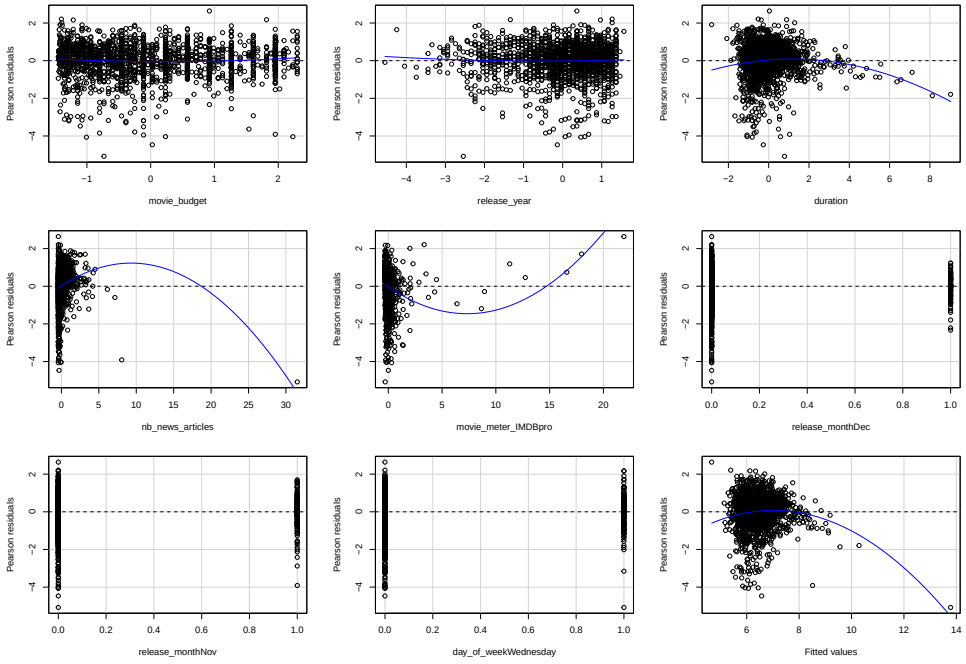

Figure 3: Residual Plots of Multivariate Model

Table 6: VIF Table

imdb_score	1.542
movie_budget	1.369
release_year	1.427
duration	1.596
nb_news_articles	1.122
actor1_star_meter	1.067
actor2_star_meter	1.205
actor3_star_meter	1.134
nb_faces	1.093
action	1.482
adventure	1.375
scifi	1.266
thriller	1.568
musical	1.096
romance	1.214
western	1.084
sport	1.131
horror	1.391
drama	1.542
war	1.127
animation	1.168
crime	1.444
maturity_ratingPG	8.547
‘maturity_ratingPG-13‘	16.277
maturity_ratingR	18.768
aspect_ratio	1.185
day_of_weekMonday	1.032
day_of_weekSaturday	1.037
day_of_weekSunday	1.034
day_of_weekThursday	1.042
day_of_weekTuesday	1.065
day_of_weekWednesday	1.166
release_monthAug	1.875
release_monthDec	1.758
release_monthFeb	1.797
release_monthJan	2.045
release_monthJul	1.806
release_monthJun	1.807
release_monthMar	1.812
release_monthMay	1.522
release_monthNov	1.883
release_monthOct	2.042
release_monthSep	1.941

Table 7: Individual Significance of Predictors

p-value	**Significance**	**Variable Names**
$p < 0.01$	Highly Significant	movie_budget, release_year, duration, nb_news_articles, movie_meter_IMDBpro, nb_faces, action, adventure, scifi, thriller, horror, drama, war, crime, release_monthDec, release_monthNov, day_of_weekWednesday
$p < 0.05$	Significant	sport, release_month_Oct, day_of_week_Saturday
$p < 0.1$	Marginally Significant	western, actor2_star_meter, release_monthJun

Table 8: Summary of Multivariate Regression[1]

	Dependent variable:
	IMDB Score
No. of Faces on the Main Poster	-0.044^{***} (0.010)
Action	-0.283^{***} (0.059)
Adventure	-0.072 (0.067)
Scifi	0.088 (0.072)
Thriller	-0.038 (0.052)
Horror	-0.340^{***} (0.072)
Drama	0.376^{***} (0.049)
War	0.140 (0.114)
Crime	0.193^{***} (0.056)
Mvoie Budget	-0.117^{***} (0.022)
Release Year	-0.143^{***} (0.022)
Duration	0.304^{***} (0.024)
Articles in the News	0.430^{***} (0.031)
2023 Ranking of Movie by IMDbPro	-0.047^{**} (0.020)
Movie released in December	0.046 (0.083)
Movie released in November	0.052 (0.079)
Movie released on a Wednesday	0.131^{*} (0.074)
Constant	6.409^{***} (0.048)
Observations	1,802
R^2	0.350
Adjusted R^2	0.344
Residual Std. Error	0.855 (df = 1784)
F Statistic	56.566^{***} (df = 17; 1784)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Tukey Test

	Test stat	Pr(>|Test stat|)
No. of Faces on the Main Poster	0.591	0.555
Action	-0.255	0.799
Adventure	0.205	0.838
Scifi	-0.297	0.766
Thriller	-0.247	0.805
Horror	-0.822	0.411
Drama	0.105	0.916
War	1.485	0.138
Crime	-0.770	0.442
Movie Budget	2.700	0.007
Release Year	0.250	0.803
Duration	-5.204	0.00000
No. of Articles in the News	-10.873	0
2023 Ranking of Movie by IMDbPro	8.044	0
Movie released in December	-0.374	0.708
Movie released in November	-0.263	0.792
Movie released on a Wednesday	-0.031	0.976
Tukey test	-9.254	0

Table 10: Summary of Multivariate Regression with degree 2[1]

	Dependent variable:
	imdb_score
nb_faces	-0.041^{***} (0.009)
action	-0.300^{***} (0.056)
adventure	-0.021 (0.064)
scifi	0.066 (0.068)
thriller	-0.044 (0.050)
horror	-0.375^{***} (0.068)
drama	0.357^{***} (0.047)
war	0.217^{**} (0.108)
crime	0.184^{***} (0.053)
poly(movie_budget, 2)1	-6.247^{***} (0.921)
poly(movie_budget, 2)2	3.216^{***} (0.823)
release_year	-0.158^{***} (0.022)
poly(duration, 2)1	12.159^{***} (0.982)
poly(duration, 2)2	-4.252^{***} (0.845)
poly(nb_news_articles, 2)1	11.439^{***} (0.868)
poly(nb_news_articles, 2)2	-8.316^{***} (0.841)
poly(movie_meter_IMDBpro, 2)1	-1.591^{*} (0.832)
poly(movie_meter_IMDBpro, 2)2	5.867^{***} (0.862)
release_monthDec	0.030 (0.078)
release_monthNov	0.079 (0.075)
day_of_weekWednesday	0.096 (0.070)
Constant	6.415^{***} (0.046)
Observations	1,802
R^2	0.418
Adjusted R^2	0.411
Residual Std. Error	0.810 (df = 1780)
F Statistic	60.923^{***} (df = 21; 1780)
Note:	*p<0.1; **p<0.05; ***p<0.01

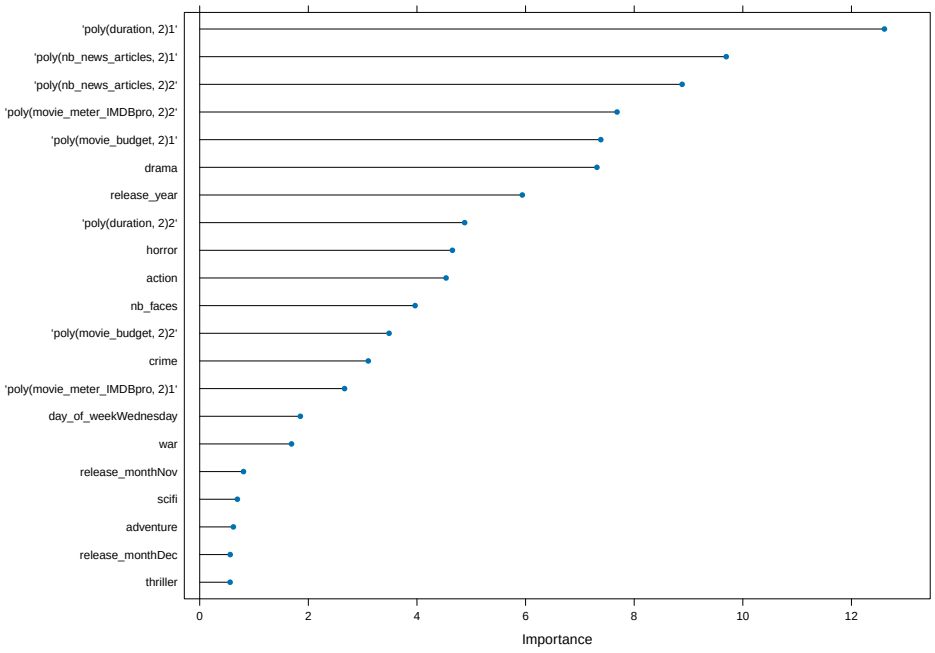

Figure 4: Variable Importance Features

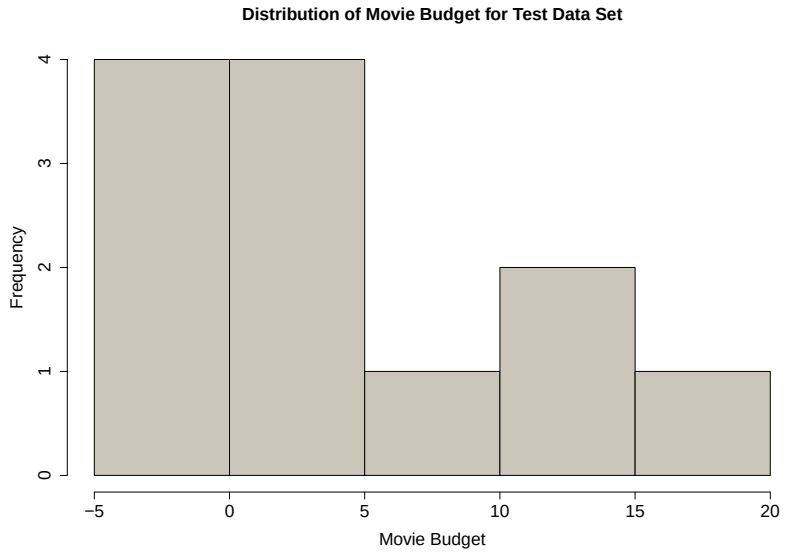

Figure 5: Distribution of Movie Budget for Test Data Set (post-standardization)
Standardizing the Movie Budget with the distribution of the test data set resulted in 3 movies with z-score greater than 10. This translates to IMDB score predictions to go beyond 10 for the 3 movies highlighted in Table 4.

References

1. Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer